

CS4220: Knowledge Discovery Methods in Bioinformatics

Course Briefing

Limsoon Wong



Recommended “Pre-requisites”

- **Completed modules on**
 - Programming
 - Algorithms
 - Basic molecular biology
 - **ST2334 Probability & Statistics**
 - **CS2220 Introduction to Computational Biology**

Objectives

- Exposure to knowledge discovery techniques
 - Enhance flexible & logical problem solving skill
 - Understand bioinformatics problems and their solution in depth
 - A modern network-based perspective
- To achieve goals above, we expose students to case studies spanning gene expression and proteomic analysis, protein functional prediction, epistatic interaction analysis, etc.

Contents of Course Overview

- **Time Table**
- **Course Syllabus**
- **Course Homepage**
- **Teaching Style**
- **Project, Assignments, Exams**
- **Readings**
- **Assessment**

- **Quick Overview of Themes and Applications of Bioinformatics**

Time Table

- **Lecture**
 - Thursday 2pm – 4pm, COM2-108
- **Tutorial**
 - Thursday 4pm – 5pm, COM2-108
- **Email**
 - wongls@comp.nus.edu.sg
- **Consultations**
 - Any time; just make appt to make sure I am in
 - Pls email my PA, tanps@comp.nus.edu.sg

Course Syllabus

- **Essence of Biostatistics**
 - Statistical estimation
 - Hypothesis testing, including
 - Non-parametric methods
- **Essence of Data Mining**
 - Clustering
 - Association rules
 - Classification
 - Class-imbalance learning
- **Gene Expression Profile Analysis**
 - Basic gene expression analysis
 - Batch effect & normalization
 - Improving reproducibility
- **Proteomic Profile Analysis**
 - Basic proteomic profile analysis
 - Improving consistency
 - Improving coverage
- **Protein Interaction Network**
 - Consistency, comprehensiveness of pathway databases
 - Integration of pathway databases
 - Reliability of PPI network
 - Identifying noise and missing edges in PPI networks
- **Protein Complex Prediction**
- **Protein Function Prediction**
- **MicroRNA and MicroRNA Target Identification**
- **Genome-Wide Association Studies**

Course Homepage

- **IVLE**
 - https://ivle.nus.edu.sg/lms/public/list_course_public.aspx?code=cs4220&acadyear=2011%2f2012
- **Lecture Slides & etc**
 - <http://www.comp.nus.edu.sg/~wongls/courses/cs4220/2012>

Teaching Style

- **Bioinformatics is a broad area**
- **Need to learn a lot of material by yourself**
 - Reading books
 - Reading papers
 - Practice on the web
- **Don't expect to be told everything**

Assignments, Project, & Exam

- **Assignments (30-40% of marks)**
 - 3 to 4 assignments
 - Some are simple programming assignments
- **Project (20-30% of marks)**
 - Based on a case study in the class
 - 8-10 pages of report / ppt slides expected
- **Exam (40% of marks)**
 - 1 final open-book exam

Be Honest

- **Exam**
 - Absence w/o good cause results in ZERO mark
 - Cheating results in ZERO mark
- **Discussion on assignments & project is allowed**
- **Blatant plagiarism is not allowed**
 - Offender gets ZERO mark for assignment or exam
 - Penalty applies to those who copied AND those who allowed their assignments to be copied

Background Readings

- **Every lecture will be accompanied by a small set of “must-read” and “good-to-read” articles**
 - The “must-read” articles are considered lecture notes and are examinable
- **For basic materials, you can read the following:**
 - Limsoon Wong, *The Practical Bioinformatician*, WSPC, 2004
 - Wing-Kin Sung, *Algorithms in Bioinformatics: A Practical Introduction*, CRC, 2010
 - These additional materials are not examinable

Related Courses

- **CS2220 Introduction to Computational Biology**
 - Understand bioinformatics problems; interpretational skills
- **CS3225 Combinatorial Methods in Bioinformatics**
- **CS4220 Knowledge Discovery Methods in Bioinformatics**
 - Clustering; classification; association rules; Mining and analysis of seq, trees, & graphs
- **CS5238 Advanced Combinatorial Methods in Bioinformatics**
 - Seq alignment, whole-genome alignment, suffix tree, seq indexing, motif finding, RNA sec struct prediction, phylogeny reconstruction
- **CS6280 Computational Systems Biology**
 - Dynamics of biochemical and signaling networks; modeling, simulating, & analyzing them
- **Etc ...**

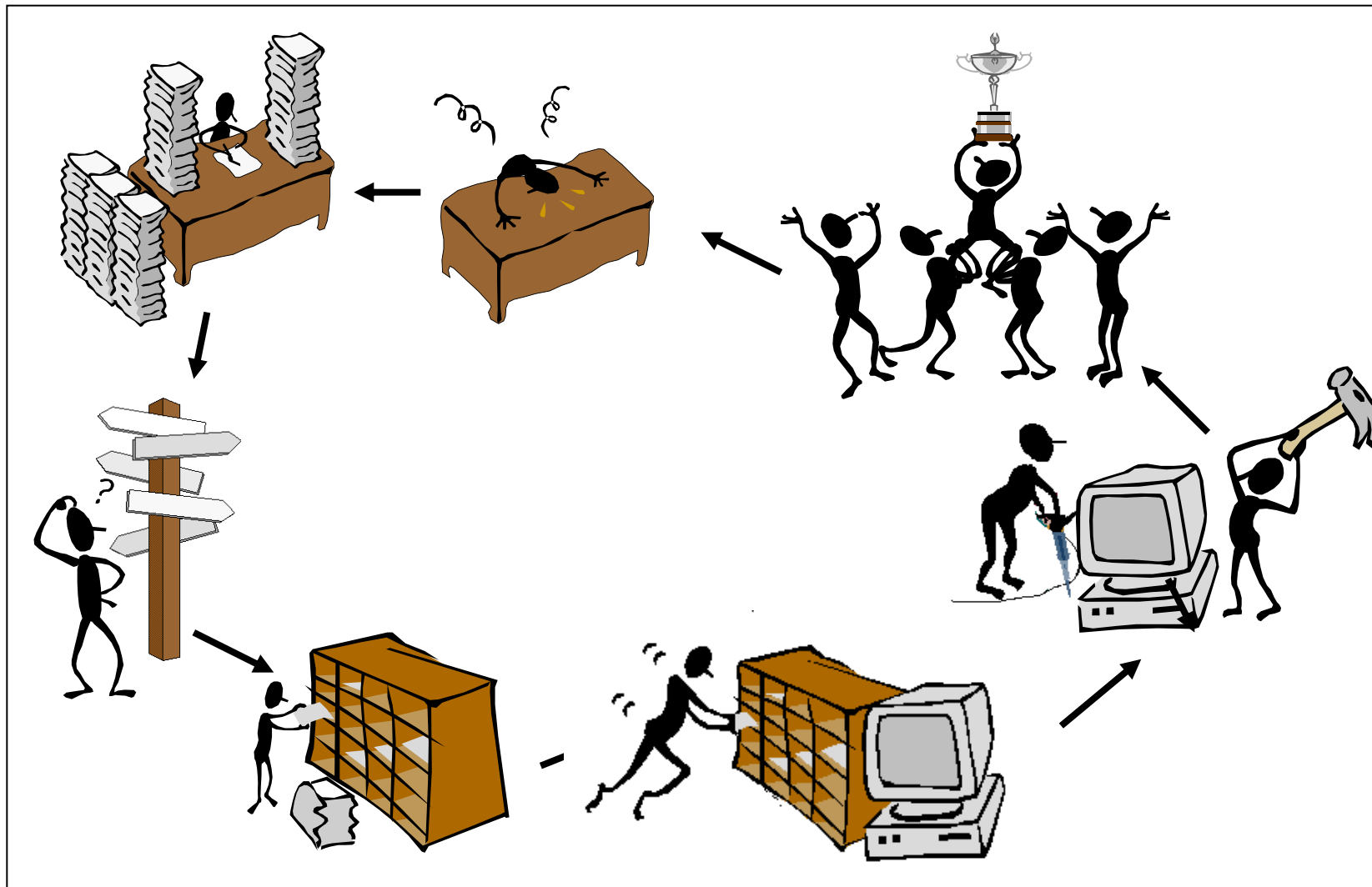
Any questions?

I hope you will enjoy this class 😊

Themes and Applications of Bioinformatics



What is Bioinformatics?



Themes of Bioinformatics

Themes of This Course

Bioinformatics involves

Data Mgmt +

Knowledge Discovery +

Sequence Analysis +

Physical Modeling + ...

Knowledge Discovery =

Statistics + Algorithms + Databases

The Promises of Bioinformatics

To the patient:

Better drug, better treatment

To the pharma:

Save time, save cost, make more \$

To the scientist:

Better science

Fulfilling the Promise via Drugs

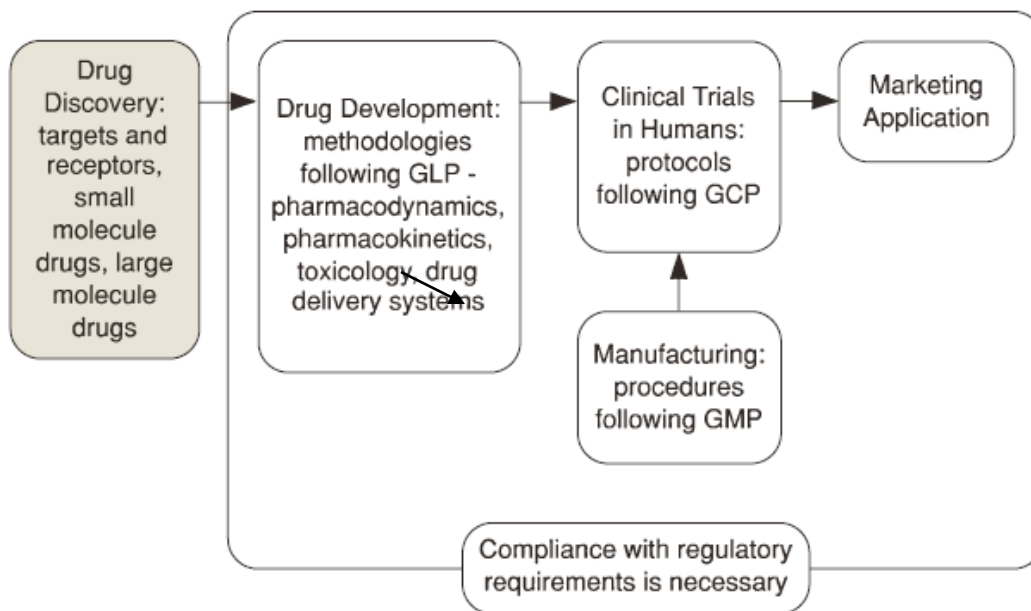


Figure from Rick Ng, *Drugs: From Discovery to Approval*

- **Bioinformatics is applicable to drug development**
- **Drug discovery: Design small molecules that bind target proteins**
 - Which proteins?
 - What should binding accomplish?
- **Biomarkers**

Pervasiveness of Bioinformatics

- **Bioinformatics is mandatory for large-scale biology**
 - e.g., High-throughput, massively-parallel measurements, or “lab on a chip” miniaturization
- **Computational data analysis is mandatory for indirect experimental methods**
 - e.g., reconstruction based on phase contrast or wave diffraction.
- **What about the rest of biology (and medicine) ?**
- **Limitless opportunities!**

Some Bioinformatics Problems

- **Biological Data Searching**
- **Biological Data Integration**
- **Gene/Promoter finding**
- **Cis-regulatory DNA**
- **Gene/Protein Network**
- **Protein/RNA Structure Prediction**
- **Evolutionary Tree reconstruction**
- **Infer Protein Function**
- **Disease Diagnosis**
- **Disease Prognosis**
- **Disease Treatment Optimization, ...**

Commonly Used Data Sources



Type of Biological Databases

- **Micro Level**

- Contain info on the composition of DNA, RNA, Protein Sequences

- **Metadata**

- Ontology
- Literature

- **Macro Level**

- Contain info on interactions
 - **Gene Expression**
 - **Metabolites**
 - **Protein-Protein Interaction**
 - **Biological Network**

Exercise: Name a protein seq db and a DNA seq db

Transcriptome Database

- Complete collection of all possible mRNAs (including splice variants) of an organism
- Regions of an organism's genome that get transcribed into messenger RNA
- Transcriptome can be extended to include all transcribed elements, including non-coding RNAs used for structural and regulatory purposes

Exercise: Name a transcriptome database

Gene Expression Databases

- **Detect what genes are being expressed or found in a cell of a tissue sample**
- **Single-gene analysis**
 - Northern Blot
 - In Situ Hybridization
 - RT-PCR
- **Many genes: High throughput arrays**
 - cDNA Microarray
 - Affymetrix GeneChip® Microarray

Exercise: Name a gene expression database

Metabolites Database

- A metabolite is an organic compound that is a starting material in, an intermediate in, or an end product of metabolism
 - Metabolites dataset are also generated from mass spectrometry which measure the mass the these simple molecules, thus allowing us to estimate what are the metabolites in a tissue
- **Starting metabolites**
 - Small, of simple structure, absorbed by the organism as food
 - E.g., vitamins and amino acids
 - **Intermediary metabolites**
 - The most common metabolites
 - May be synthesized from other metabolites, or broken down into simpler compounds, often with the release of chemical energy
 - E.g., glucose
 - **End products of metabolism**
 - Final result of the breakdown of other metabolites
 - Excreted from the organism without further change
 - E.g., urea, carbon dioxide

Protein-Protein Interaction Databases

- **Proteins are true workhorses**
 - Lots of cell's activities are performed thru PPI, e.g., message passing, gene regulation, etc.
- **Methods for generating PPI db**
 - biochemical purifications, Y2H, synthetic lethals, in silico predictions, mRNA-co-expression
- **Function of a protein depends on proteins it interacts with**
- **Contain many false positives & false negatives**

Exercise: Name a PPI database

Introductory References

- **S.K. Ng, “Molecular Biology for the Practical Bioinformatician”, *The Practical Bioinformatician*, Chapter 1, pages 1-30, WSPC, 2004**
- **DOE HGP Primer,**
http://www.ornl.gov/sci/techresources/Human_Genome/publicat/primer/index.shtml
- **Lots of useful videos,**
http://www.as.wvu.edu/~dray/Bio_219.html
- **Materials from CS2220,**
<http://www.comp.nus.edu.sg/~wongls/courses/cs2220/2011/>